63-3-4

AD NO. ______ 403502

IA FILE COPY

446 500

(6) See next p

(7) NA

(8) NL

(9) NA

(10) See next p

(11) See next p

(12) 18 p

(13) NA

(13) See next p

(16) NA

(17) NA

(18) NA

(19) NA

(21) NA

(14) CDRP-182-114

Report Number

403 502

United States Atomic Energy Commission
Division of Technical Information

$1.60

OTI-131 (May 1962)

USAEC Division of Technical Information Extension, Oak Ridge, Tennessee

SPEARMAN'S FOOTRULE --
AN ALTERNATIVE RANK STATISTIC.

By
D. C. Kleinecke
H. K. Ury
L. F. Wagner

ABSTRACT

A known but neglected rank statistic -- Spearman's Footrule -- is examined. Formulas are given for its mean and variance and the covariance with better known rank statistics. A very thorough numerical description is given, including the exact sampling distributions for up to ten ranks when all permutations are equally likely and approximate results up to twenty. The statistic appears to be asymptotically normal but this has not been proven. There are numerous applications for this statistic in non-parametric testing, but they are not discussed in this report.

Civil Defense Research Project
Institute of Engineering Research
College of Engineering
University of California - Berkeley

CDRP-182-114
~~November 1962~~

Office of Civil Defense
Department of Defense
Contract ~~No.~~ CD-SR-58-40

A

INTRODUCTION

The Civil Defense Research Project in its studies of fallout prediction has been using a model of wind behavior which treats the entire set of winds aloft as a stationary stochastic process. In actual computations with this model it has been assumed that the intercomponent covariances are essentially zero when the time between winds is greater than three days. This assumption is based on the observed fact that the covariances do tend fairly rapidly toward zero. No test has actually been made, however, of the significance, in a statistical sense, of the covariances at time differences of more than three days.

It was suggested that, in order to improve the accuracy of wind statistics, the covariances for periods of up to a week or so might be used. This would involve a considerable increase in computing time and, clearly, it is not desirable to extend the time range beyond the point where the covariances are insignificant. Thus, the problem can be restated somewhat as follows: determine the longest time range for which the winds aloft are statistically dependent in a significant fashion.

The rather unusual statistical problem just formulated has been discussed occasionally in the literature. The only approaches which combine practical applicability with generality seem to be those which are based ultimately on rank comparison tests. Rank statistics are not very powerful but they are absolutely nonparametric and do not depend on assumptions about the underlying distribution of the winds. The authors feel that the usual rank statistics are not completely satisfactory for the problem at hand and the unusual statistic (Spearman's footrule) discussed below was proposed as a substitute. Due to circumstances beyond the control of the authors, the program outlined in the paragraphs above was not carried out and the statistical dependence of the winds was never investigated. The rank statistic was studied in some detail, however; the results of this study are being reported here as a technical contribution to statistics.

SPEARMAN'S FOOTRULE

Rank statistics seem to be a seldom used side branch of classical statistics, in spite of their respectable antiquity. The literature on the subject has been surveyed thoroughly by M. G. Kendall (ref. 1). (We must, of course, point out that not all of his interpretations can be accepted, especially those involving correlation values considerably less than one.) Nevertheless, the present increasing interest in nonparametric methods makes it likely that an ever-increasing usage of rank statistics can be safely predicted.

Kendall introduced (to all intents and purposes) and studied in great detail a statistic which he called τ. Kendall's τ is clearly the best single rank statistic. There is, however, a need for alternative rank statistics to support τ and for use in those situations where τ is, for some reason, a priori unacceptable. The best known rank statistic, Spearman's ρ, is not really adequate for these purposes because it is asymptotically equivalent to τ (the correlation between τ and ρ goes to one as the number of ranks goes to infinity); it seems that ρ is useful only as a convenient approximation to τ.

There is one more rank statistic which was proposed by Spearman (ref. 2). This is defined below and will be denoted by δ. Kendall dismisses it with the comment that there are analytical difficulties in dealing with the sampling distribution. It is true that the sampling distribution is not easy to handle but it is not hopeless as will be shown below, and there are compensations.

Suppose there are n objects which have a natural ordering that can be symbolised by identifying the objects with the first n integers. Consider a rearrangement of these n objects; this can be symbolised by a permutation p of the first n integers. The three rank statistics are:

$$\tau = \Sigma_{ij}\, \mathrm{sgn}(i-j)\mathrm{sgn}(p_i-p_j)$$

$$\rho = \Sigma_i (p_i-i)^2$$

$$\delta = \Sigma_i |p_i-i|$$

(actually Kendall uses τ and ρ for versions of these statistics normalised to vary between minus one and plus one; since there does not seem to be any advantage to imitating a correlation and there is always the possibility of misinterpretation if the statistics are normalised, this will not be done in what follows). In spite of appearances ρ resembles τ more than δ; for the details of this see Kendall (ref. 1).

The main contents of this paper are a set of tables giving the exact sampling distribution, assuming all permutations are equally likely, of the statistic δ for the cases n = 1,2,3,4,5,6,7,8,9,10 and Monte Carlo estimates of the distribution for the cases n = 11,12,13,14,15,20. This represents a more complete numerical description of δ than is available for either τ or ρ. Figures are included which show histograms of the distribution for these cases; it should be clear from these figures that the distribution is smooth, unimodal, and rapidly tends to the normal with increasing n in spite of some skewing. Although it has not been proved mathematically that δ is asymptotically normally distributed (the proof given in Kendall, ref. 1, does not generalize to this case), it seems like a reasonable working hypothesis. The smooth lines on the histograms show the normal approximations obtained by using the same mean and variance as of δ itself.

The trend toward normality seems to be rapid enough to permit use of the normal approximation in any of the cases $n > 10$ where the exact distribution is not known, provided, of course, that no attempt is made to evaluate the extreme tails of the distribution. For $n \geq 14$ the approximation, used

with the standard (Cochran) continuity correction, appears to be accurate to at least two places except for the extreme upper tail of the distribution; the accuracy was studied at the 10%, 5%, 2½% and 1% significance levels. At the upper tail the approximation will obviously result in overestimates. Here, surprisingly good results have been obtained by using it without the continuity correction, but more work needs to be done on this.

The sampling distribution under the assumption that all permutations are equally likely can be validly applied only to questions involving the hypothesis that two (or more) rankings are statistically independent. This means that in general only the lower tail is of interest in applications to testing. Most of the asymmetry appears to fall on the opposite tail, which again encourages the use of the normal approximation. Kendall (ref. 1) has commented that the distribution of δ is not as spread out as that of ρ and implies that this means that δ is therefore, a priori, less useful than ρ or τ. This conclusion is not necessary, however, because the fact that δ has fewer different values only means that tests based on δ can be applied at fewer different levels of significance. This is not an operationally important restriction because, for $n \geq 5$, tests based on δ can be applied at a set of values which adequately covers any range of practically interesting significance levels. For two-sided tests this will hold for $n \geq 8$.

As was implied above, the general expression for the moments of δ, even asymptotically, has not been found, and neither has the characteristic function or any generating function -- although it is clear that an expression in terms of permanents like that given by Kendall (ref. 3) can be obtained even if it cannot be used effectively. The first two moments, however, have been calculated, and they are given below; the joint moments with ρ and τ have also been calculated; similar results involving ρ and τ alone are available in Kendall (ref. 1).

$$\text{Mean }(\delta) = \frac{1}{3}(n+1)(n-1)$$

$$\text{Variance }(\delta) = \frac{1}{15}(n+1)(2n^2+7)$$

$$\text{Mean }(\tau) = 0$$

$$\text{Variance }(\tau) = \frac{2}{9}n\,(n-1)(2n+5)$$

$$\text{Covariance }(\delta,\tau) = -\frac{2}{15}(n+1)(n^2+1)$$

$$\text{Mean }(\rho) = \frac{1}{6}n(n+1)(n-1)$$

$$\text{Variance }(\rho) = \frac{1}{36}n^2(n+1)^2(n-1)$$

$$\text{Covariance }(\delta,\rho) = \frac{1}{30}n\,(n+1)(n^2+1)$$

$$\text{Covariance }(\tau,\rho) = -\frac{1}{9}n(n+1)^2(n-1)$$

Expressions for the correlations are obtained immediately and need not be reproduced here. It is also clear that, asymptotically as n goes to infinity

$$\text{Correlation }(\delta,\tau) \to -\frac{3}{\sqrt{10}} = -0.95$$

$$\text{Correlation }(\rho,\tau) \to -1$$

naturally, the correlation δ and ρ is essentially the same as that of δ and τ. This demonstrates the fact mentioned above that ρ and τ are asymptotically equivalent.

All of the relationships given above are obtained in the same general manner which can be illustrated by one example -- the variance of δ which illustrates the full range of complexity. First, the expectation of δ^2 is to be computed. If $\mathcal{E}$ is the expectation operator, then

$$\mathcal{E}\delta^2 = \Sigma_{ik}\, \mathcal{E}(|p_i-i|\,|p_k-k|)$$

$$= \Sigma_i\, \mathcal{E}|p_i-i|^2 + \Sigma_{i\neq k}\, \mathcal{E}(|p_i-i|\,|p_k-k|)$$

$$= \Sigma_i \frac{1}{n} \Sigma_j |j-i|^2 + \Sigma_{i\neq k} \frac{1}{n(n-1)} \Sigma_{j\neq \ell} |j-i|\,|\ell-k|$$

$$= \frac{1}{n-1} \Sigma_{ij}(j-i)^2 - \frac{2}{n(n-1)} \Sigma_i(\Sigma_j|j-i|)^2 + \frac{1}{n(n-1)} (\Sigma_{ij}|j-i|)^2$$

The third term can be evaluated at once in terms of the mean of δ which can be assumed as known; the first term can be evaluated by straightforward summation; the second term is evaluated as follows:

$$\Sigma_i(\Sigma_j|j-i|)^2 = \Sigma_i(\Sigma_{j=1}^{i-1} (i-j) + \Sigma_{j=i+1}^{n}(j-i))^2$$

$$= \Sigma_i(i^2 - (n+1)i + \tfrac{1}{2}n\,(n+1))^2$$

$$= \frac{1}{60}n\,(n-1)(n+1)(7n^2-8)$$

thus

$$\mathcal{E}\delta^2 = \frac{1}{6}n^2(n+1) - \frac{1}{30}(n+1)(7n^2-8) + \frac{1}{9}n\,(n-1)(n+1)^2$$

$$= \frac{1}{90}(n+1)(5n^3-3n^2-5n+12)$$

and the variance is obtained at once by subtracting the square of the mean.

The method just outlined for the evaluation of the variance can be extended, with some difficulty, to higher order central moments. This has been done only for the third and fourth moments of δ. The results are complicated and will not be given here. The third moment goes asymptotically to zero (as is required for the normal limit to be valid) as $n^{-\frac{1}{2}}$ and the fourth moment is also asymptotically correct.

The tables given below for the exact sampling distribution of δ for $n = 1,2,\ldots,10$ were obtained by a SAP-coded program for the IBM-704 computer. The entire set, which involved the generation and evaluation of more than $3\frac{1}{2}$ million permutations, was obtained in less than one hour's running time. The Monte Carlo approximations for the cases $n = 11,12,13,14,15,20$ were obtained from samples of 100,000 randomly chosen permutations. The moments of the sample set were computed and compared with the theoretical values given by the formulas quoted above. A sample was rejected if either its mean or variance fell outside the 70% confidence interval for the corresponding theoretical moment. This was the shortest confidence interval consistent with holding the expected cost of computer time below a specified limit. In case of rejection of the initial sample, an attempt was made to pool it with a subsequent one, thereby utilizing the larger number of trials. This resulted in the use of the combined first and second samples for $n = 13$ and 15 and of the pooled first and third samples for $n = 14$. For $n = 11$, 12 and 20, the initial sample was accepted.

The user of the tables for $n = 11$, 12, 13, 14, 15 and 20 should be aware that the total number of trials is either 100,000 or 200,000 as given in the tables and the tabulated number is the number of times each value of δ was observed. If the exact number of times each value occurs when all permutations are considered is required, it can be approximated by renormalizing to $n!$ instead of the number of trials. Note that this will give zero as the approximate number for several values on the lower tail which clearly do actually occur although rarely. On the other hand, all of the sampled cases assume the maximum possible value -- the greatest integer in $n^2/2$ -- a relatively large number of times.

REFERENCES

1. Kendall, M. G. Rank Correlation Methods, Hafner Publishing Co., 1955.

2. Spearman, C. "A Footrule for Measuring Correlation," British Jour. Psych. 2 (1906) 89.

3. Kendall, M. G. The Advanced Theory of Statistics, Vol. 1, 1954 (5th ed.).

ACKNOWLEDGMENT

The authors wish to express thanks to Lila Richardson for typing the manuscript and to Richard Moriarty for the drawings.

EXACT DISTRIBUTION OF δ

δ	n = 1	2	3	4	5	6	7	8	9	10	δ
0	1	1	1	1	1	1	1	1	1	1	0
2		1	2	3	4	5	6	7	8	9	2
4			3	7	12	18	25	33	42	52	4
6				9	24	46	76	115	164	224	6
8				4	35	93	187	327	524	790	8
10					24	137	366	765	1,400	2,350	10
12					20	148	591	1,523	3,226	6,072	12
14						136	744	2,553	6,436	13,768	14
16						100	884	3,696	11,323	27,821	16
18						36	832	4,852	17,640	50,461	18
20							716	5,708	25,472	83,420	20
22							360	5,892	33,280	127,840	22
24							252	5,452	40,520	182,256	24
26								4,212	44,240	242,272	26
28								2,844	45,512	301,648	28
30								1,764	40,608	350,864	30
32								576	35,496	382,576	32
34									25,632	389,232	34
36									18,108	373,536	36
38									8,064	332,640	38
40									5,184	273,060	40
42										208,548	42
44										136,512	44
46										81,792	46
48										46,656	48
50										14,400	50
Totals:	1	2	6	24	120	720	5,040	40,320	362,880	3,628,800	

FREQUENCIES OF δ OBTAINED BY MONTE CARLO METHODS

δ	n = 11	n = 12	n = 13	n = 14	n = 15	δ
0	0	0	0	0	0	0
2	0	0	0	0	0	2
4	0	0	0	0	0	4
6	0	1	0	0	0	6
8	2	0	0	0	0	8
10	12	1	0	0	0	10
12	24	1	2	0	0	12
14	61	12	3	0	0	14
16	146	34	8	1	0	16
18	298	46	18	0	0	18
20	594	123	32	6	0	20
22	958	217	81	14	0	22
24	1,578	365	166	25	1	24
26	2,482	612	276	58	11	26
28	3,432	1,022	477	96	16	28
30	4,644	1,538	766	167	27	30
32	5,941	2,176	1,239	278	55	32
34	6,943	3,093	1,800	450	83	34
36	8,065	3,845	2,568	696	164	36
38	8,966	4,765	3,538	1,002	221	38
40	9,397	5,865	4,914	1,503	353	40
42	9,411	6,752	6,194	1,961	552	42
44	8,911	7,484	7,810	2,768	790	44
46	7,681	8,063	9,436	3,757	1,108	46
48	6,582	8,345	11,043	4,790	1,564	48
50	4,988	8,416	12,491	5,966	2,143	50
52	3,874	7,869	13,410	6,995	2,683	52
54	2,355	7,101	14,462	8,526	3,499	54
56	1,612	6,000	14,881	9,953	4,211	56

(continued on following page)

FREQUENCIES OF δ OBTAINED BY MONTE CARLO METHODS (cont.)

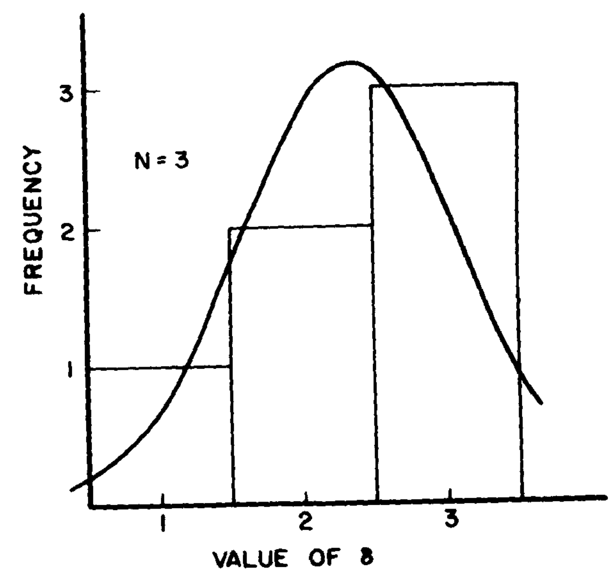
δ	n = 11	n = 12	n = 13	n = 14	n = 15	δ
58	646	5,113	14,668	10,847	5,241	58
60	397	3,928	14,473	12,127	6,359	60
62		2,839	13,177	13,037	7,323	62
64		1,965	11,910	13,247	8,486	64
66		1,198	10,396	13,427	9,514	66
68		718	8,759	13,043	10,510	68
70		372	6,873	12,651	11,067	70
72		101	5,033	11,761	11,928	72
74			3,618	10,670	12,146	74
76			2,546	9,395	12,324	76
78			1,399	8,126	11,984	78
80			917	6,422	11,486	80
82			411	5,129	10,891	82
84			205	3,834	9,938	84
86				2,879	8,938	86
88				1,861	7,796	88
90				1,264	6,556	90
92				677	5,489	92
94				375	4,387	94
96				176	3,343	96
98				40	2,350	98
100					1,701	100
102					1,166	102
104					747	104
106					445	106
108					248	108
110					101	110
112					55	112
Total no. of trials	100,000	100,000	200,000	200,000	200,000	

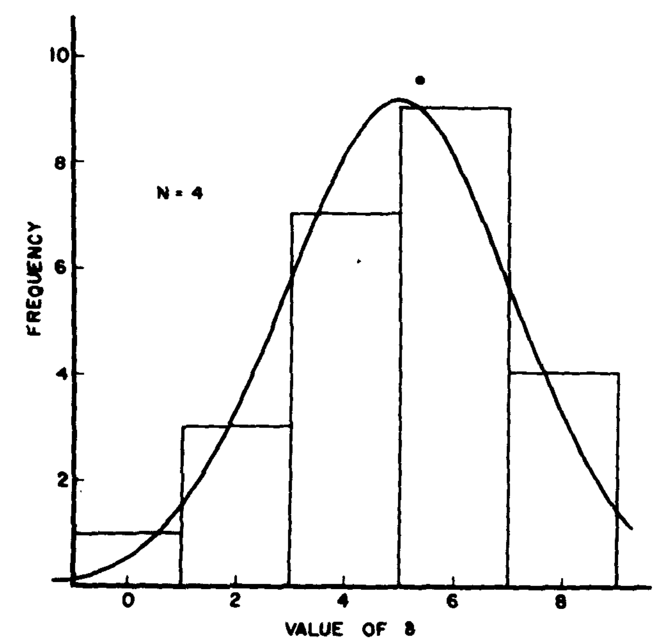
FREQUENCIES OF δ OBTAINED BY MONTE CARLO METHODS (cont.)

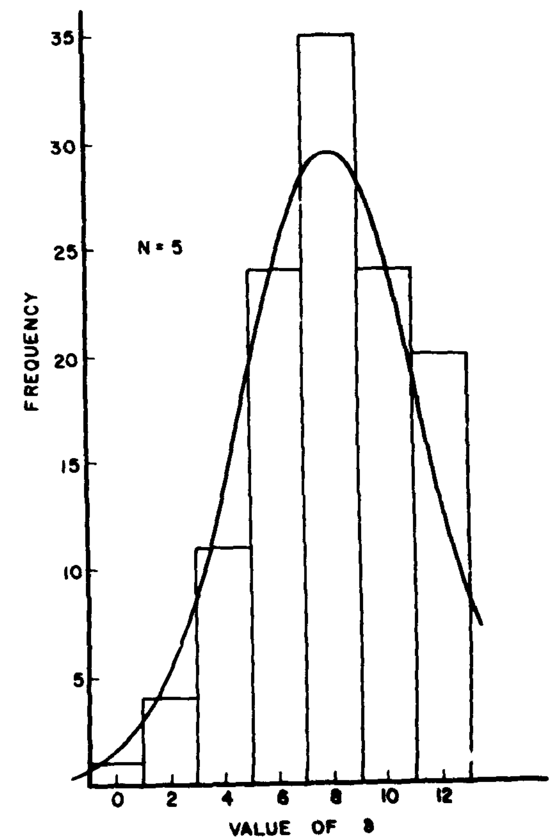
$n = 20$

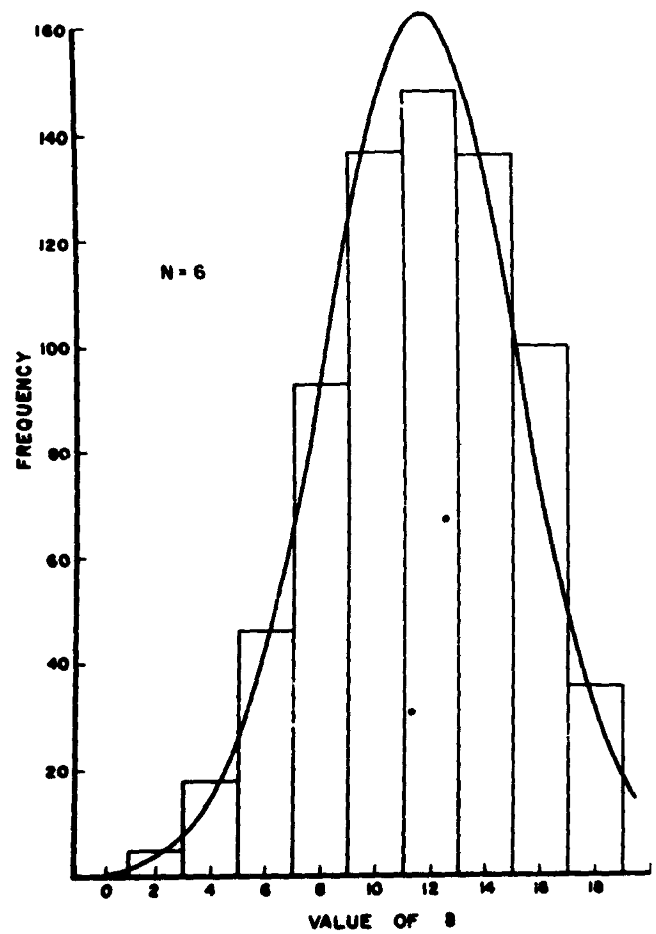
δ		δ		δ		δ	
0	0						
2	0	52	0	102	1,191	152	2,750
4	0	54	0	104	1,382	154	2,376
6	0	56	0	106	1,511	156	2,075
8	0	58	3	108	1,775	158	1,869
10	0	60	3	110	1,921	160	1,670
12	0	62	6	112	2,214	162	1,482
14	0	64	10	114	2,344	164	1,161
16	0	66	7	116	2,682	166	990
18	0	68	18	118	2,976	168	855
20	0	70	24	120	3,215	170	723
22	0	72	33	122	3,301	172	537
24	0	74	36	124	3,646	174	431
26	0	76	65	126	3,661	176	324
28	0	78	86	128	3,887	178	237
30	0	80	104	130	4,023	180	191
32	0	82	157	132	3,988	182	137
34	0	84	192	134	3,973	184	82
36	0	86	238	136	3,977	186	63
38	0	88	286	138	4,015	188	35
40	0	90	411	140	3,964	190	20
42	0	92	492	142	3,770	192	16
44	0	94	602	144	3,562	194	10
46	1	96	736	146	3,433	196	2
48	0	98	847	148	3,227	198	1
50	0	100	1,001	150	2,965	200	2

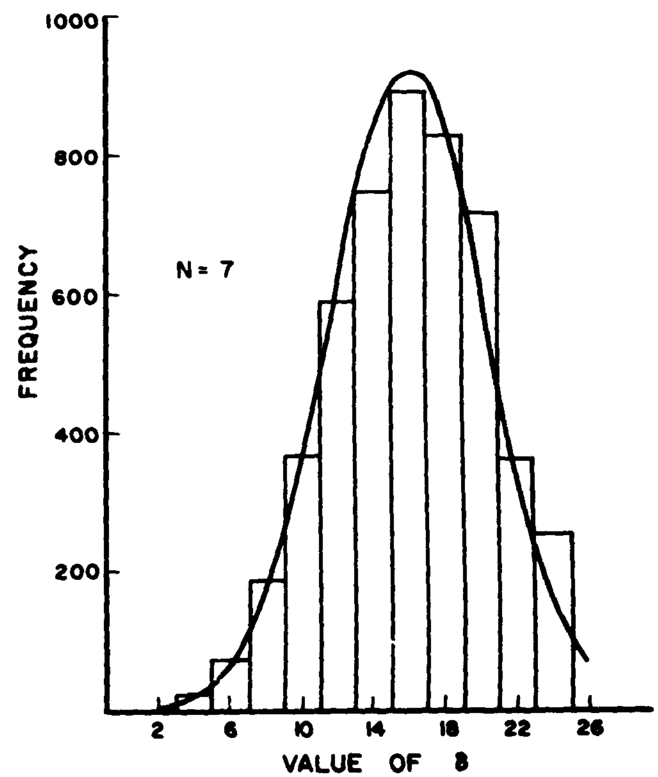
Total number of trials: 100,000

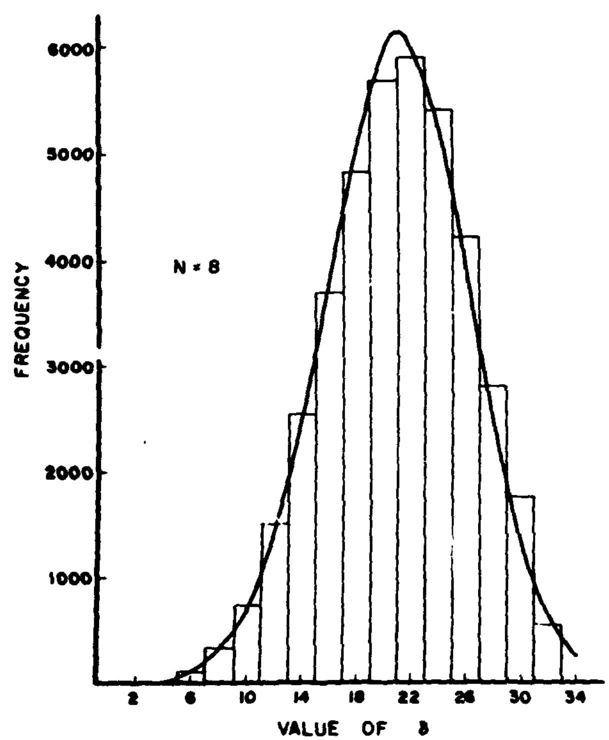


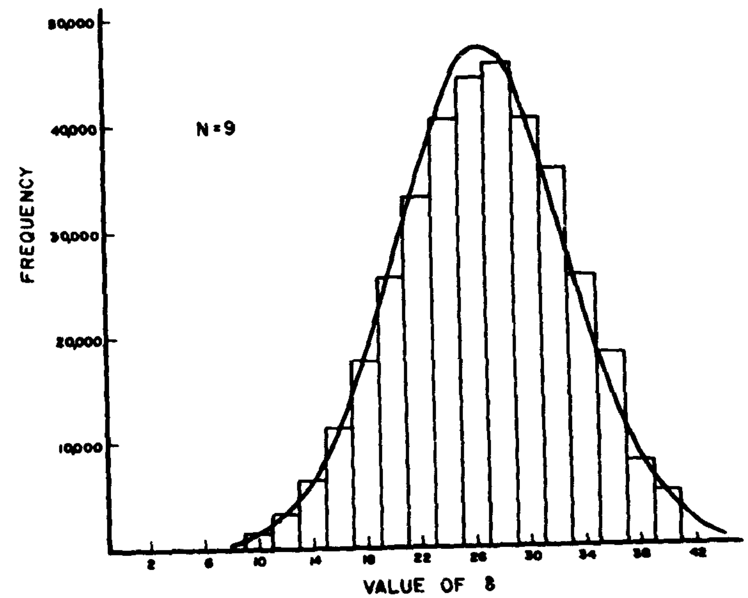

N = 4
FREQUENCY
10
8
6
4
2
0
2
4
6
8
VALUE OF 8

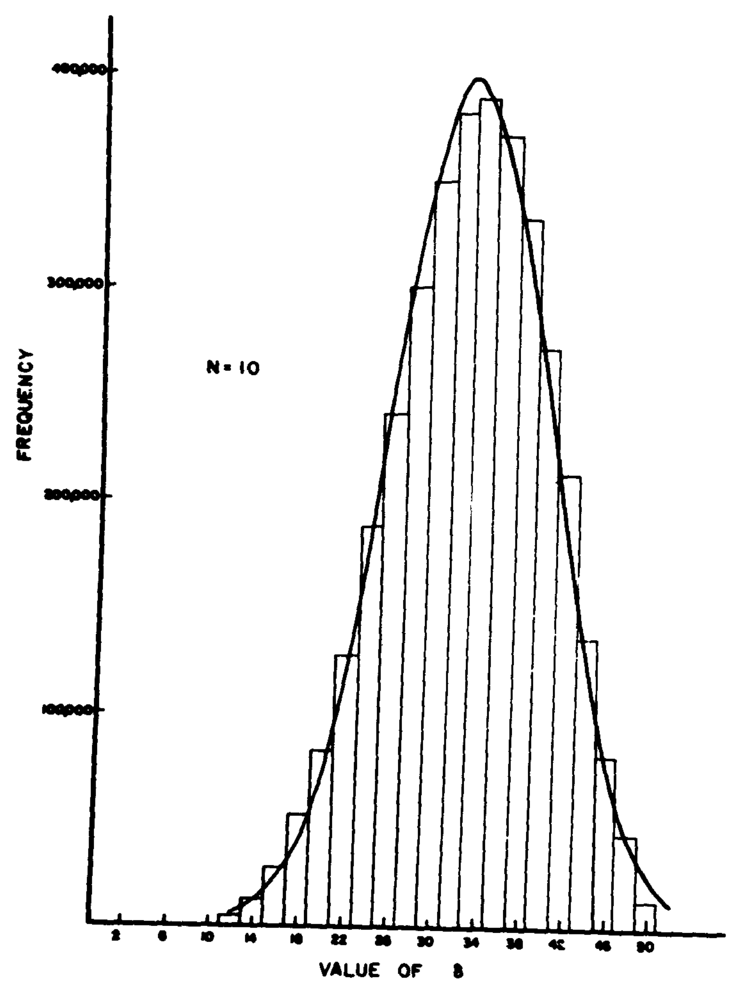

N = 5
FREQUENCY
35
30
25
20
15
10
5
0
2
4
6
8
10
12
VALUE OF 8

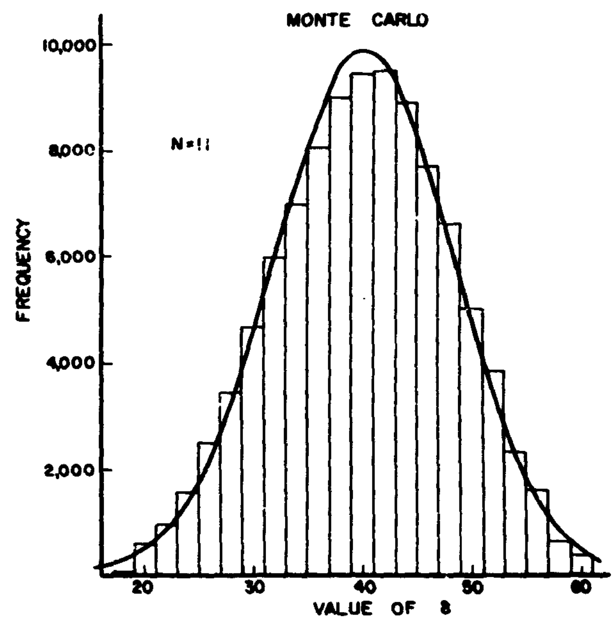

N = 6
FREQUENCY
160
140
120
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
VALUE OF S

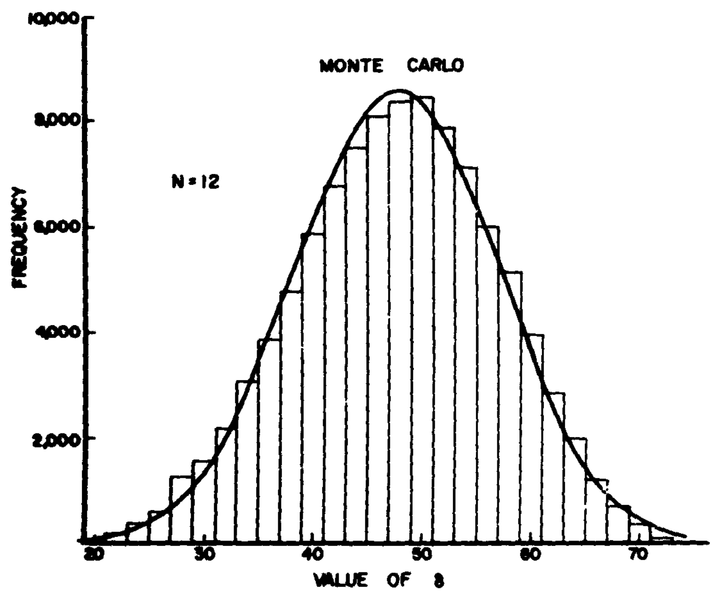

N = 7
FREQUENCY
1000
800
600
400
200
2
6
10
14
18
22
26
VALUE OF S

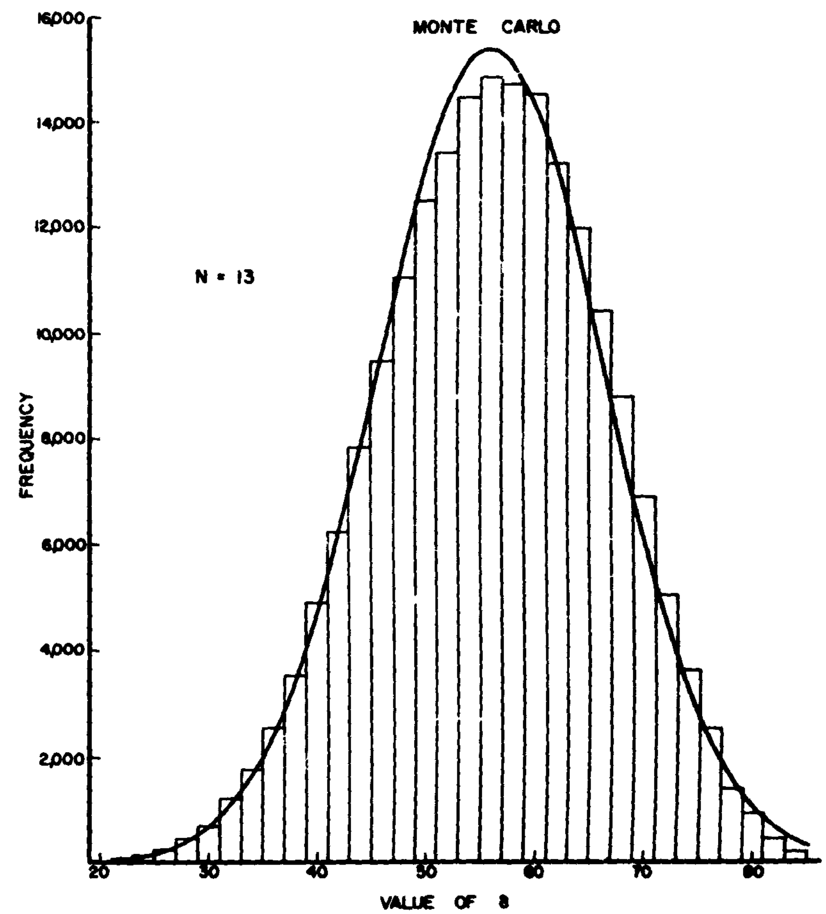

N = 8
FREQUENCY
VALUE OF δ

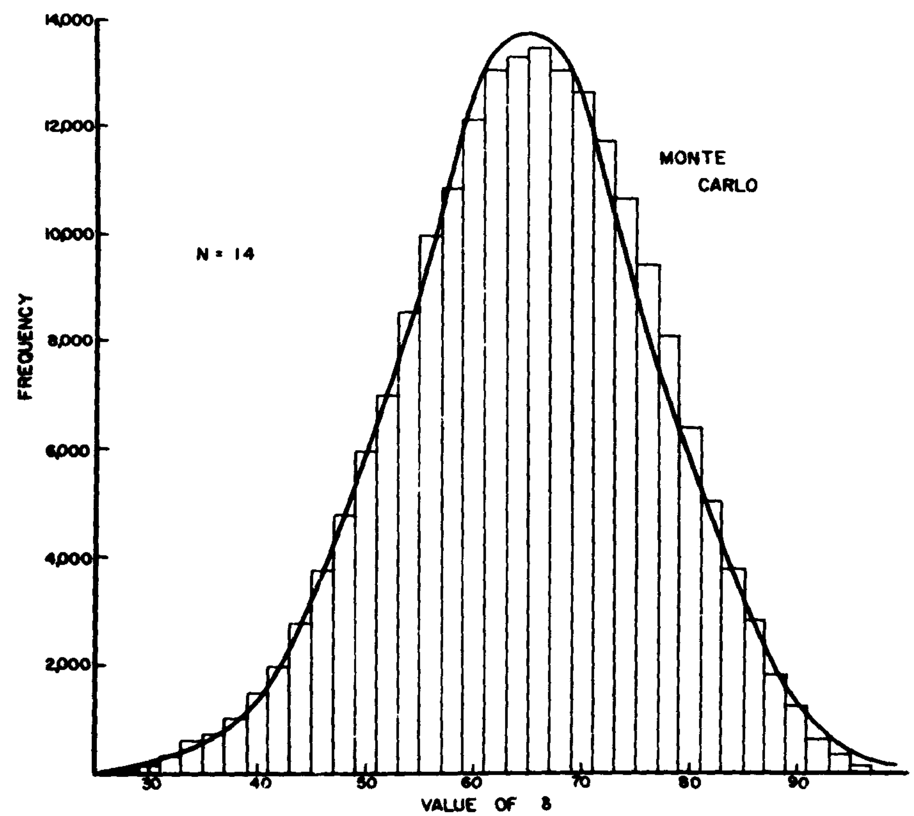

N = 9
FREQUENCY
VALUE OF δ


FREQUENCY
400,000
300,000
200,000
100,000
N = 10
2
6
10
14
18
22
26
30
34
38
42
46
50
VALUE OF S

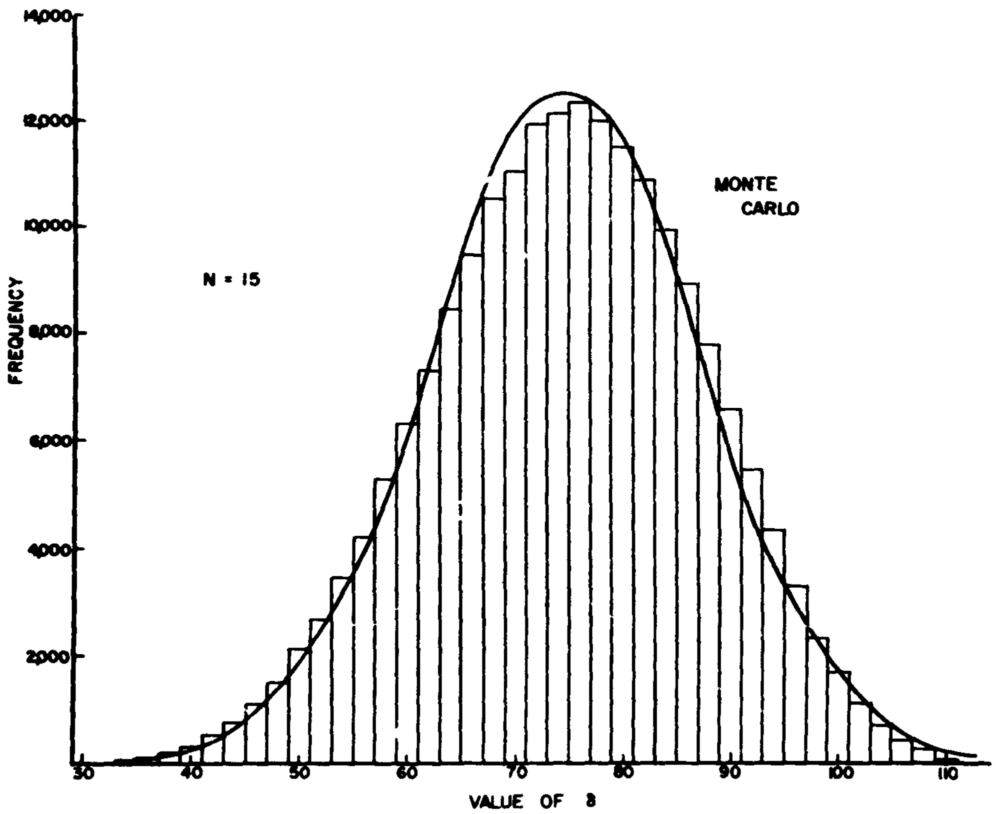

MONTE CARLO
FREQUENCY
10,000
8,000
6,000
4,000
2,000
N = 11
20
30
40
50
60
VALUE OF S


MONTE CARLO
N = 12
FREQUENCY
10,000
8,000
6,000
4,000
2,000
20
30
40
50
60
70
VALUE OF 8

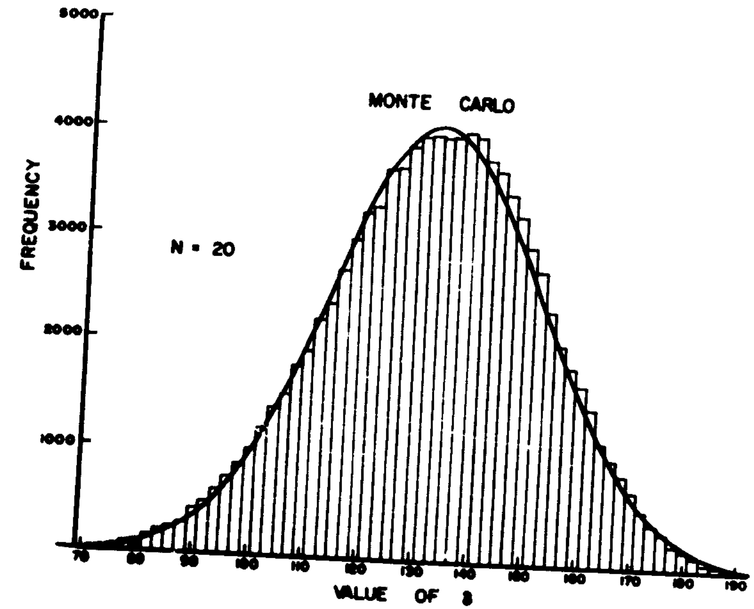

MONTE CARLO
N = 13
FREQUENCY
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
20
30
40
50
60
70
80
VALUE OF 8



14,000
12,000
10,000
8,000
6,000
4,000
2,000
FREQUENCY
N = 14
MONTE
CARLO
30
40
50
60
70
80
90
VALUE OF S



14,000
12,000
10,000
8,000
6,000
4,000
2,000
FREQUENCY
30
40
50
60
70
80
90
100
110
VALUE OF 8
N = 15
MONTE
CARLO


MONTE CARLO
N = 20
FREQUENCY
5000
4000
3000
2000
1000
70
80
90
100
110
120
130
140
150
160
170
180
190
VALUE OF 8

END